Different scales of *function* for a given bio molecule X

- **Chemical / physical** (microscopic scale): binds another molecule, catalyzes a molecular reaction, etc.
- **Biological** (macroscopic scale): leads to a phenomenologic / phenotypic transformation
- All scales in between the above (mesoscopic)
- X may have >1 function, across / within these scales
- A general / naïve test for *function*: Perturb X in native system and observe what happens at all scales
Eg. mutation (frameshift, mis-sense / non-synonymous) of methyl-CpG-binding protein 2 (MECP2, Xq28) -> Rett syndrome, a progressive, X dominant neurologic developmental disorder. Phenotype incl. autism, dystonia, short, etc. Typ. fatal in males (major encephalopathy). Females -> somatic X mosaicism.

- **MECP2 chemical function**: binds methylated DNA -> repress transcription from methylated gene promoters
- **MECP2 biological function**: embryonic development
- Mutation (truncating frameshift, mis-sense) of cyclin dependent kinase-like 5 (CDKL5, Xp22) leads to almost similar phenotype. CDKL5 chemical functions: ATP binding, protein serine / threonine kinase activity, nucleotide binding

2 archetypal questions in functional genomics

- What function does a given molecule X have in a specific biological system - state?
- Which molecules (their interactions) “associate” with / “underwrite” a given biological system - state?
Module 4: Functional Genomics (FG) lecture 1 outline

- FG definition. 2 iconic studies

- Review basic functional concepts: gene, expression, epigenetics, uni-directional transfer of \textit{genetic} information ("central dogma")

- Survey of parallel high-throughput bio quantification technologies
  - Scalable detection principles: sequencing short oligomers, nucleotide complementarity
  - Representative technologies: SAGE, microarrays. Assumptions / Pros / Cons.
  - Technical generalizations

- Transcriptome studies: basic caveats / assumptions

- Shift in perspective / way to think about biological problems
Functional genomics is the deconstruction of the genome to

- Ascribe *function* to genes & non (protein) coding genomic elements / NCGE's – different levels of *function*
- Characterize interactions between genes & NCGE's

- Using the vast genomic evolutionary, sequential, structural, expression information.
- Leverage on synergy of diverse data modalities

*Proxies for “Function” & Raw genomic info*

Graph adapted from Figure 1 of:
Classical lymphoma histopath classification unify different morphologic subtypes into 1 group, eg. diffuse large B cell lymphoma (DLBCL)

Transcriptome-scale profiles of 96 lymphatic malignancies (mostly DLBCL, CLL, FL) and normal tissue. All DLBCL patient de novo and biopsy samples obtained pre-treatment. Questions:

- Identify distinct molecular portraits for DLBCL malignancies
- Identify DLBCL malignancy subtypes new to current classification system
- Relate each malignancy to distinct stage of normal B cell development

Hierarchical clustering using full transcriptome features reveal heterogeneity within DLBCL subgroup

Hierarchical clustering of entire dataset using full transcriptome features – reveals ordered heterogeneity among samples.

GC B-cells genes relevant, specific to B-cell development, use this subset of features to re-cluster DLBCL.

Figure removed due to copyright reasons. Please see figure 1 from:
Hierarchical clustering of DLBCL using B-cell develop relevant subsets of transcriptome

Feature-induced *Regularities* within DLBCL

Reveal 2 distinct molecular subgroups with significantly different survival outcomes

Figure removed due to copyright reasons.
Please see figures 3, 5a, and 5c from:
“It is important to note that considerable gene expression heterogeneity exist within each subgroup, and no single gene in either of these clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy.”

Talking points from study #1:

- New diagnostic subcategory of DLBCL missed by morphologic + immunohistochemical (with a few markers) analyses marred by irreproducibility.
- New prognostic tool and corresponding therapeutic opportunities.
- Hypothesis generation for basic biology of DLBCL. [1] Now we have a clearer sense of the granularity of DLBCL cases. [2] Mechanistic differences between these 2 DLBCL subgroups?
- Genomic data can be fruitfully exploited without mechanistic functional assignment
Reverse engineer the reaction network architecture of early glycolysis from metabolite in(2) / out(8) put time series – using time-lagged correlation + multi-dimensional scaling

Classically determined pathway of early glycolysis

In / Out put time series

FG: Iconic study #2 (Arkin et al, Science 1997)

CMC/MDS predicted pathway of early glycolysis

Causality arrow inferred from temporal ordering

Classically determined pathway of early glycolysis

“[Metabolite input] ranges represent the extreme “physiological” concentration attained by these species.”

Talking points from study #2:

- Not all metabolites known to be involved / produced in the process were measured
- Certain interactions between species that were measured were not resolved
- Analysis is sensitive to initial conditions, eg. pH, temperature, initial inflow species concentrations.
DNA is a physical molecule. **Genome** = total cellular DNA. What is a gene?

- 1854-65 *“Unit factors” of inheritance*, Gregor Mendel (Brno)
- 1869 *Nucleic acid / DNA isolated*, Johann Miescher (Tübingen)
- 1952 *DNA (not protein) might be genetic material / agent*, Alfred Hershey & Martha Chase (Cold Spring Harbor)
- 1953 *DNA is genetic material / agent (structurally makes sense)*, James Watson, Francis Crick & Rosalind Franklin (Cambridge, UK)
- 2005 Lolle et al. (Nature March 23 issue) epigenetic (non-Mendelian) recovery of HOTHEAD gene in cress

Definition of a gene (NCBI)

- A fundamental physical and **functional** unit of heredity that is a DNA sequence located on a specific site on a chromosome which encodes a specific functional product (eg. RNA, protein)
Example zoom into a contiguous subset of the genome

What's the non-gene stuff in the genome? (eukarya)

- **Genes** (~1.5% genome. Eg. protein coding exons), gene-related DNA (~36% genome. Eg. non-coding introns – eukarya, pseudogenes), intergenic DNA (~62.5% genome. Eg. microsatellites, genome-wide repeats). Coding = transmission into mRNA.

- **Genome-wide repeats**. E.g., transposons, long/short interspersed nuclear elements

Eukaryote vs. Prokaryote (operons, no introns) genomes

- C
FG: Review concept of gene expression

- **Definition of gene expression (NCBI)**
  - *The process by which information encoded in a gene is transcribed into RNA, and then typically into protein.*

- **Gene expression is a function of cellular *state***
  - Time – eg, developmental stage
  - Space – eg, organ. tissue
  - Other state variables – eg, disease, environmental cues

- **Transcriptome** = all mRNA present in a cell at a particular state

![Gene Expression (eukarya)](chart)

**Alternative Splicing**

![Alternative Splicing](chart)

Different isoforms -> different function, i.e., different proteins translated.

Figure by MIT OCW.
Recall different levels of *function* (chemical, biological, etc). (protein)

Coding and non-coding RNA categories:

- **Messenger RNA** = protein coding transcripts, typically high degradation rate
- **Transfer RNA** = transfer aa to polypeptide chain during translation
- **Ribosomal RNA** = primary (structural) constituent of ribosomes
- **Small nuclear RNA** = RNA splicing, telomere maintenance, form snRNProteins
- **Small nucleolar RNA** = chemical modification (eg. methylation) of rRNA
- **Guide RNA** = RNA editing in protozoa
- **Micro RNA** = RNA interference at post/pre-transcription

Figure removed due to copyright reasons. Please see figure 1.14 in:
Epigenetic definition

- Heritable change in gene function *without* nuclear DNA sequence change
- Selective gene in/activation within an organism. Eg. X mosaicism, imprinting, cell fate determination

A key process: methylation (H -> CH$_3$) of DNA or protein

- **DNA methylation** typically on CpG sites, catalyzed by DNA methyltransferases (DNMT's)
- CpG sites (5'CG3') < expected prob. given uniform random base dist. due to DNMT's.
- CpG islands: >200bp, GC% > 50%, obs/exp CpG ratio > 0.6, high density near eukarya 5' gene promoters. Tend to be un-CH$_3$.
- +CH$_3$ @ CpG islands modulates corresponding gene's expression.
- Protein methylation typically on arginine/R, lysine/K, catalyzed by protein CH$_3$-transf. Esp. histones (post translational modif.) modulate local gene activity
- Bisulfite sequencing assay: bisulfite -NH$_2$ but not 5-methylcytosine

![Chemical structures indicating DNA methylation](image)
Functional expression of methylation

- Embryonic development: 1 to 8 cell stage, eukaryotic genome de-CH$_3$. 8-cell to morula (~32 cell blastomere), de novo +CH$_3$. By blastula stage, +CH$_3$ complete. DNA CH$_3$-transf knockout embryos die at morula stage.

- Environment factors (cellular stress -> polyamines) modulate CH$_3$ pattern postnatal development

- Imprinting: Prader-Willi/Angelman (chr 15q11.2-11.3)

- CH$_3$ state as cancer, neoplastic biomarkers

<table>
<thead>
<tr>
<th>Gene</th>
<th>+CH$_3$ of CpG islands in promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td>v-abl Abelson murine leukemia viral oncogene homolog 1</td>
<td>50-100% chronic myelogenous leukemia</td>
</tr>
<tr>
<td>chondroitin sulfate proteoglycan 2</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>deleted in bladder cancer chromosome region candidate 1</td>
<td>50% bladder cancer</td>
</tr>
<tr>
<td>endothelin receptor B</td>
<td>60-70% prostate cancer</td>
</tr>
<tr>
<td>Wilm's tumor 1</td>
<td>90% breast cancer</td>
</tr>
<tr>
<td>pi-class glutathione S-transferase</td>
<td>80-100% prostate cancer</td>
</tr>
</tbody>
</table>

- Other epigenetic disorders

  - Mitochondrial diseases (matrilineal) – typically metabolic
Original statement, “central dogma” (CD) of molecular biology

- The [CD] deals with detailed residue-by-residue transfer of sequential information ... such information cannot be transferred from protein to either protein or nucleic acids. [Crick, 1958]

Over-simplified (mis-interpreted) CD

- DNA to RNA to Protein

- (faux) Exceptions: Retroviruses (by reverse transcriptase), DNA modifying proteins,
Recall, **transcriptome** = all mRNA present in a cell at a particular state, organism-space-time specific

- **Identification / characterization**
  - Genomic libraries: DNA fragments of (near) total genome @ *specific state*
  - cDNA libraries: mRNA fragments (no intron) -> cDNA fragments -> sequence -> expressed sequence tags (EST's), GenBank ID#
  - 1 gene “covered” by >1 EST’s. Eg. human genome >4M EST’s, ~30K genes
  - Screen EST’s -> EST’s assoc with a particular gene form a Unigene cluster
  - Differential comparison between cDNA libraries: Binary analysis (present/absent). $H_0$: # of seq for a given gene X is the same in two libraries. Prob test: Fisher exact.
    Limitations: sequencing error + depth, tissue of origin contamination, library construction bias

---

**UniGene – EST map not well-defined**

- Gene A1 -> EST B1, EST B3, EST B6
- Gene A2
- Gene A3 -> EST C1
### FG: Concept of Unigene cluster

- **Human FoxP2 gene has 52 EST's in it's Unigene cluster (Hs.282787)**

<table>
<thead>
<tr>
<th>GenBank ID</th>
<th>Description</th>
<th>Tissue of Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF700673.1</td>
<td>Clone IMAGE:4285527, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>T97069.1</td>
<td>Clone IMAGE:121181, 5' read</td>
<td>mixed</td>
</tr>
<tr>
<td>T96957.1</td>
<td>Clone IMAGE:121181, 3' read</td>
<td>mixed</td>
</tr>
<tr>
<td>BUS521502.1</td>
<td>Clone IMAGE:6527367, 5' read</td>
<td>uterus</td>
</tr>
<tr>
<td>BQ948273.1</td>
<td>Clone IMAGE:6473507, 5' read</td>
<td>uterus</td>
</tr>
<tr>
<td>AL711700.1</td>
<td>Clone DKFzP686E0284, 5' read</td>
<td>muscle</td>
</tr>
<tr>
<td>BM725479.1</td>
<td>Clone UI-E-EJ0-aie-p-18-0-UI, 5' read</td>
<td>other</td>
</tr>
<tr>
<td>BM701645.1</td>
<td>Clone UI-E-EJ0-ahl-h-24-0-UI, 5' read</td>
<td>other</td>
</tr>
<tr>
<td>BJ752226.1</td>
<td>Clone IMAGE:5192788, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>N31133.1</td>
<td>Clone IMAGE:265380, 5' read</td>
<td>skin</td>
</tr>
<tr>
<td>N21118.1</td>
<td>Clone IMAGE:265380, 3' read</td>
<td>skin</td>
</tr>
<tr>
<td>DN990126.1</td>
<td>Clone TC100653, 5' read</td>
<td>Whole brain</td>
</tr>
<tr>
<td>AV658847.1</td>
<td>Clone GLCFQG08, 3' read</td>
<td>liver</td>
</tr>
<tr>
<td>AV658824.1</td>
<td>Clone GLCFQE09, 3' read</td>
<td>liver</td>
</tr>
<tr>
<td>CV573230.1</td>
<td>Clone od33g10, 5' read</td>
<td>lung</td>
</tr>
<tr>
<td>CR738014.1</td>
<td>Clone IMAP998F084735; IMAGE:1929991, 5' read</td>
<td>embryonal kidney</td>
</tr>
<tr>
<td>BP871788.1</td>
<td>Clone HKR01979</td>
<td>mammary gland</td>
</tr>
<tr>
<td>BE068078.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637513.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637512.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637511.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637510.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637509.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637508.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>BX481950.1</td>
<td>Clone DKFzP686D03228, 5' read</td>
<td>muscle</td>
</tr>
<tr>
<td>CD001942.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410738.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410682.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410681.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>BX280996.1</td>
<td>Clone IMAP998G13581; IMAGE:265380</td>
<td>skin</td>
</tr>
<tr>
<td>CB118125.1</td>
<td>Clone B1T694954-5-A03, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>T06261.1</td>
<td>Clone HFBRD02</td>
<td>brain</td>
</tr>
<tr>
<td>AI459612.1</td>
<td>Clone IMAGE:2152081, 3' read</td>
<td>colon</td>
</tr>
<tr>
<td>AI624789.1</td>
<td>Clone IMAGE:2234155, 3' read</td>
<td>uterus</td>
</tr>
<tr>
<td>AI798932.1</td>
<td>Clone IMAGE:2348762, 3' read</td>
<td>mixed</td>
</tr>
<tr>
<td>CK430225.1</td>
<td>Clone oj46f12, 5' read</td>
<td>eye</td>
</tr>
<tr>
<td>CV569620.1</td>
<td>Clone od07e09, 5' read</td>
<td>eye</td>
</tr>
<tr>
<td>BF678535.1</td>
<td>Clone IMAGE:4250207, 5' read</td>
<td>prostate</td>
</tr>
<tr>
<td>BG722650.1</td>
<td>Clone IMAGE:4826916, 5' read</td>
<td>testis</td>
</tr>
<tr>
<td>BI495413.1</td>
<td>Clone IMAGE:2539657, 3' read</td>
<td>other</td>
</tr>
</tbody>
</table>

### Human Unigene clusters

**Oct 2005**

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>155,852</td>
<td>mRNAs</td>
</tr>
<tr>
<td>4,514</td>
<td>Models</td>
</tr>
<tr>
<td>48,605</td>
<td>HTC</td>
</tr>
<tr>
<td>1,574,398</td>
<td>EST, 3'reads</td>
</tr>
<tr>
<td>2,431,310</td>
<td>EST, 5'reads</td>
</tr>
<tr>
<td>1,114,365</td>
<td>EST, other/unknown</td>
</tr>
<tr>
<td>5,329,044</td>
<td>Total sequences in clusters</td>
</tr>
</tbody>
</table>

**Approx # human genes ~30K**

FG: Parallel high-throughput transcriptome profiling technologies

- Low throughput (1 RNA species at a time): northern blot, RT-PCR

- 2 scalable principles for detecting / quantifying gene transcription products, and their representative technologies
  - Sequencing short representative sub-sequence (unsupervised): serial analysis of gene expression (SAGE). Sequence frequency $\propto$ abundance
  
  - Nucleotide base pair complementation of short representative sub-sequences (supervised): cDNA / RNA microarray. Fluorescent intensity $\propto$ abundance

  - Unsupervised = the universe of measurable entities is not constrained by the assaying platform. However, mapping these entities to known RNA species depends upon reference sequence library.

  - Supervised = universe of RNA species which are measurable is constrained by the assaying platform

- Pre-assay steps: From a biological system at specific state -> extract mRNA -> form cDNA (more stable), fragmented. Amplify? Bias towards 3’ end targets, other non-linear artifact.
FG: SAGE (sequencing short representative sub-sequences)

- **SAGE**
  - Have a SAGE library: bijective map between SAGE tags and genes / EST's
  - Obtain mRNA to construct corresp cDNA.
  - From each cDNA transcript, cut a short sequence tag (SAGE tag) 10-14 bps from a specific position (3'-end typically) that will uniquely identify that transcript.
  - Tags have uniform length.
  - Concatenate all tags into one concatamer -> clone -> sequence.
  - # of times a particular tag observed = expression level of particular gene

- Details @ www.bioteach.ubc.ca/MolecularBiology/PainlessGeneExpressionProfiling
SAGE example result: 3 transcript types relative to a SAGE library

Table removed due to copyright reasons.
Please see www.embl-heidelberg.de/info/sage.
Pros

- Discover new genes, or old genes with new role (function / tissue-time specificity)
- Abundance of a transcript = simple counting

Cons

- Tag specificity. Short SAGE tag size may lead to identification problems. 1 tag mapping to >1 genes is a problem.
- Restriction enzyme action variability. Each SAGE tag must have constant length, otherwise problems arise in sequencing concatamer. Restriction enzyme may not yield tags of uniform length. Not all mRNA species have the same enzyme recognition sequence, plus temperature dependent.
- What is appropriate Control / Reference system for comparison? This is a more general problem that we will see as we progress in functional genomics.
Definition RNA / cDNA microarray (chip)

- Single-stranded DNA (gene / EST sub-sequences) anchored at one end on a substrate, e.g. gridded array or bead surface. Different species placed on separate grid coordinates / beads. ssDNA fragment (called probes), not entire gene sequence is placed. Why?

- Evolved from southern blots (DNA). Exploits parallelism

- Mechanistic principle: Nucleotide complementarity A ↔ T, G ↔ C

- ssDNA on chip will hybridize to complementary strand in solution (cDNA's derived from a biological system, called targets). Complementary strand is fluorescent labeled.

- Basic assumption: Fluorescence is proportional to RNA abundance (thus gene expression level)
2 categories of microarrays (by manufacturing process)

- **Spotted**: (Pat Brown, Stanford). Robot attaches prepared ssDNA probes ~$10^{2-3}$ bp long on substrate. Customizable $\rightarrow$ heterogeneous (noisy)

- **Oligonucleotide**: (e.g., Affymetrix). Photolithography. Typically standardized manufacturing and shorter (relative to spotted microarrays) length oligos placed.

Figure removed due to copyright reasons.
Please see:
*Nature Genetics Supplement* 21, no. 1 (January 1999).

Figure from:
2 categories of microarrays (by usage principle): 2 channel vs. 1 channel

- **2 channel**
  - Paired experiment design
  - Can be treated as 1 channel if common control is used for all chip experiments
  - Internal (measurement, possibly biological) control / reference for fluorescence

---

**FG: Microarray categories – 2 channel**

**Control**

1. Competitive, 2-channel

2. cDNA target

3. cDNA probe

**Test**

4. cDNA probe

5. Imaging

6. Competitive, 2-channel

Courtesy of Jeremy Buhler. Used with permission.
2 categories of microarrays (by usage principle): 2 channel vs. 1 channel

- **1 channel**
  - Different internal control housekeeping genes

**Diagram:**
1. Biotin Labeled cRNA
2. Oligonucleotide microarray
3. Hybridized Array

**Additional Components:**
- SAPE: Streptavidin-phycoerythrin
FG: Generic microarray experiment stages

- Generic stages of a microarray experiment
  - Experimental design involving biological system under investigation. Replicates – both measurement / technical and biological
  - RNA and target preparation: Extract mRNA. Convert (to ss cDNA typically). Label with fluorescence.
  - Probe hybridization.
  - Fluorescence image analysis
  - **Microarray data analysis** (post image) – one lecture onto itself

Figures removed due to copyright reasons. Please see:
Microarray oligo probe design technical issues

- 3'-end target amplification bias (not strictly microarray problem). Assess by 3'-end probe-to-target / 5'-end probe-to-target intensity ratios of housekeeping probes eg. Gapdh, β-actin. Non-linear effect (with respect to diff RNA species for fixed time interval).
- CG% content of probes: C ↔ G (3 H-bonds) vs. A ↔ T (2 H-bonds) -> diff bonding energy -> diff hybridization rate. Non-linear (wrt. diff RNA species for fixed time interval)

Cons: Probe specificity. Cross RNA species hybridization, promiscuous probes

General caveats with transcriptome profiling studies

- Non-uniform RNA degradation – pre-assay step
- “Noise”: Measurement / technical and biological variation. Choice of a Reference system. This is a more general problem.
- Assumption: Central Dogma holds (mRNA ∝ Protein). Bio-process of interest engages transcriptome machinery and state is characterized by transcriptome profile
- Averaging / pooling of RNA across heterogeneous cell populations
Generalization of chip parallelism / complementarity principle

- Protein microarrays. Identify protein targets, e.g., other proteins (protein-protein interaction), mRNA, other bio-active small molecules.

- Tissue microarrays. Paraffin blocks of distinct biological tissue cores. Simultaneous histologic analysis, immunohistochemical (protein) & in situ (mRNA) analyses.

- Reverse transfection microarrays. cDNA probes on grid with a cell culture on top. Cells assimilate probes.
FG: Parallel high-throughput tech changes our perspectives / questions?

- 2 views qualitatively-diff afforded by these technologies

  - **View 1**: *Whole = Sum of individual parts*. Purely an efficient way to screen many many molecules. Multi-plexing classical assaying techniques eg. northern blot

  - **View 2**: *Whole > Sum of individual parts*. As above, plus unraveling intrinsic regularities (correlations) between measured molecules. Eg. below, do G1 or G2 intensities alone distinguish between disease groups?
... mathematical reformulation of biological problems involving multi-variables (microarrays). Next class.